DOCUMENT RESUME

ED 286 911 TM 870 543

AUTHOR Chang, Lin; Becker, Betsy Jane
TITLE A Comparison of Three Integrative Review Methods: Different Methods, Different Findings?
PUB DATE Apr 87
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Comparative Analysis; *Effect Size; Higher Education; High Schools; *Hypothesis Testing; *Meta Analysis; Natural Sciences; Research Design; Sex Differences; Statistical Analysis; *Statistical Bias; Statistical Significance; Statistical Studies
IDENTIFIERS *Science Achievement

ABSTRACT

Data drawn from 30 journal articles and ERIC documents reporting on gender differences in natural science achievement were re-examined. Three meta-analysis methods were used: (1) vote counts and vote-counting estimation procedures; (2) tests of combined significance; and (3) analyses of effect sizes. The three methods produced seemingly contradictory conclusions which were explained in terms of differences in the hypotheses tested by the methods, as well as the statistical properties of the methods. Effect-size analyses were found to be more informative for the study of single outcome constructs, whereas combined significance tests with or without effect-size analyses did not provide more information. Sex differences were not evident in many subject areas. However, males had higher achievement than females in physics by one-half of a standard deviation and in biology by about one-sixth of a standard deviation. (MGD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286911

A Comparison of Three Integrative Review Methods: Different Methods, Different Findings?

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. Chang
B. J. Becker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Lin Chang and Betsy Jane Becker

Michigan State University

Paper presented at the annual meeting of the American Educational Research Association.

Washington D. C., April, 1987

TM 870 843

A Comparison of Three Integrative Review Methods: Different Methods, Different Findings?

Lin Chang and Betsy Jane Becker

Michigan State University

Abstract

Comparisons are made of the hypotheses and conclusions appropriate to three different meta-analysis methodologies: vote counts and vote-counting estimation procedures, tests of combined significance, and analyses of effect sizes. Also, the statistical properties of related estimates and the statistical power of the hypothesis tests are discussed. Such comparisons both facilitate choosing among the methods and illustrate the strengths and weaknesses of each. Application of the three methodologies is illustrated using data from a synthesis of gender differences in science achievement. Seemingly contradictory conclusions from the three methods can be explained in terms of differences in the hypotheses tested by the methods and in the statistical properties of the methods. Analyses revealed that the magnitudes of gender differences in science achievement varied according to the subject-matter under study. Effect-size analyses were more informative than the other two approaches. Recommendations for use of the three approaches conclude the paper.

Introduction

In the past decade interest has grown in methods of synthesizing and analyzing the results of related research reports in the social science. Many researchers, confronted by a multitude of study results, have turned to quantification as a mean of handling the volume of information.

By now a great variety of statistical analyses is available for the quantitative synthesis of research results. Prospective reviewers must be able to consider the available methods and select a method (or methods) which test interesting hypotheses, support reasonably specific and informative conclusions, and have good statistical properties.

In this paper we compare three of the available approaches to the statistical analysis of related research results. We examine tests and estimates of effect magnitudes based on vote counting, tests of combined significance, and effect-size analyses. Comparisons of these methods are useful for a variety of reasons. They illustrate concretely the differences between hypotheses and conclusions for competing meta-analysis methods. The kinds of conclusions that can be justified should be one of the primary concerns of a reviewer selecting an analysis method. Also our comparisons will illustrate how differences in the statistical properties of three meta-analysis methods may lead to apparently different conclusions about a research domain. An understanding of such differences can help researchers as readers to better judge the conclusions drawn in meta-analyses of interest.

Our comparisons are based on empirical data -- a "case study" of the the use of the three methods. It is impossible to tell which of such data analyses are "correct" (unless simulated data are studied) because any real data will represent a sample with unknown population parameters. Power computations and examinations of the statistical properties of the competing

analyses can indicate how likely the analyses would be to suggest "correct" decisions under hypothetical circumstances. Such information must augment any empirical comparisons of the outcomes of the analyses. Consideration of agreement and contradictions among conclusions, in light of the strengths and weaknesses of the different methods, can increase our understanding of the functioning of the different analyses.

This paper contains four main sections. In section one the three approaches are introduced, with the hypotheses and conclusions that may be justified on the basis of each approach. Second, the literature concerning the statistical properties of the three approaches is discussed. Third, an example based on studies of gender differences in science achievement (Steinkamp & Maehr, 1983, 1984; Becker & Chang, 1986) illustrates differences in results and inferences for the three approaches. Finally, results based on the three methods are related to statistical properties of the methods and recommendations are given for the use of the methods.

Hypotheses and Conclusions for the Three Approaches

We first introduce some notation to represent parameters for the study outcomes under investigation. Mean differences are often used in comparing average or typical performance for two groups. Consider a series of k studies, the population effect size (or standardized mean difference) δ for the two groups within study i ($i = 1, \ldots, k$) is defined as

$$\delta_i = (\mu_{1i} - \mu_{2i})/\sigma_i,$$

where the μ_{1i} and μ_{2i} are the populations means in the ith study for groups 1 and 2 respectively, and σ_i is the common population standard deviation for study i. Various approaches have been used to estimate the population effect size (e.g. Glass, 1976; Hedges, 1981; Kraemer, 1983); differences in estimation procedures are not crucial to the issues we discuss here. Other

parameters such as population proportions and correlations may be investigated using the methods we consider below; we focus on effect-size analyses primarily because they have been so widely applied.

Vote-counting Methods

The traditional vote-count procedure has been done two ways: one is to observe the proportions of positive and negative results and the other is to count results which are significant in either direction and which are nonsignificant. In either case, the category with the highest proportion of votes, or which receives more than a specified proportion of votes (e.g., 50 percent of the votes) is considered to be supported by the data. Vote-counting methods are useful because they can be applied when only the directions and significance of study results are reported.

Hedges and Olkin (1980) noted that with the traditional vote-count the meta-analyst is testing the hypothesis

$$H_0 : \delta = 0, \qquad (1)$$

where the δ is a single population effect size assumed to be common to all studies. When the procedure of observing only the directions of the results is used, the vote-count procedure is analogous to the sign test using the assumption that the chance of obtaining a positive result is 0.5. We note that the vote-count does not always differentiate between situations in which all effects are zero and those in which the average population effect size is zero and is represented by a balance of positive and negative results.

Modified vote-count procedures (Hedges, 1986; Hedges & Olkin, 1980, 1985) provide a maximum likelihood estimate of a single overall (or average) effect size (denoted here and below as $\hat{\delta}$) on the basis of counts of study outcomes. Using either a significance test or a confidence interval based

on that estimate, the reviewer can again test the hypothesis that $\delta = 0$ shown in Model 1.

The procedures for testing and estimation based on vote counts provided by Hedges (1986) do not require the assumption that all studies share a single population effect size δ. Thus these newest procedures enable the meta-analyst to examine another hypothesis about the effect sizes δ_1 through δ_k. This test concerns the consistency (homogeneity) of the population effects. The null model is

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_k = \delta, \qquad (2)$$

versus the alternative that at least one of the effect sizes differs from the hypothesized common effect size δ. When the null Model 2 is rejected, the reviewer can conclude that some studies are from populations which have either larger or smaller population effect sizes than the other populations.

Combined Significance Tests

Combined significance tests are summaries computed from the sample probabilities associated with independent tests of similar hypotheses (e.g., Rosenthal, 1978). These summaries have been extensively studied and used by statisticians for over fifty years (e.g., Fisher, 1932; Tippett, 1931).

Becker (in press) indicated that the statistical null hypothesis for combined significance tests is always the composite hypothesis of no effects in all of the studies under review. When the measures of study outcomes are the effect sizes δ_1 through δ_k, the null hypothesis examined by the combined significance tests is

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_k = 0. \qquad (3)$$

The hypothesis in Model 3 is a composite of the hypotheses in Models 1 and 2. Like the modified vote-counting procedures, such tests examine the assumption that all studies share a common effect size.

When the combined significance tests are based on probability values computed from directional tests (each examining the alternative that $\delta_i > 0$), the alternative model for the combined significance tests is

$$H_1 : \delta_i \geq 0, \text{ for } i = 1, \ldots, k, \text{ and} \qquad (4)$$

$$\text{at least one } \delta_i > 0.$$

When the null hypothesis is rejected on the basis of the directional probability values, the meta-analyst can conclude that "at least one" population effect size is nonzero and positive. Rejection of the null model thus can be associated with many different patterns of effect sizes (or other study results), and the substantive interpretations of these patterns vary greatly for all the possible unspecified alternative hypotheses. All of the studies may share a common positive effect size, or perhaps only one study has a positive population effect size. Alternatively some population effects may be positive while others are negative.

With an analysis based solely on tests of combined significance, one cannot distinguish between these different patterns of population effect sizes. Becker (in press) illustrated this with examples from three meta-analyses in which significant probability-value summaries were associated with three very different configurations of sample effect sizes. Therefore the interpretations and the inferences based on rejecting the null hypothesis for tests of combined significance are not precise even when identical population parameters are under study.

Effect-size Analyses

The composite hypothesis tested by combined significance tests can also be tested through effect-size analyses, but it typically is examined in two steps. First, one tests the hypothesis that the studies share a single population effect size. The statistical test for the homogeneity of effect

sizes is a test of the hypothesis given in Model 2 above versus the alternative that at least one of the effect sizes differs from the hypothesized common effect size δ. The computational procedures for the homogeneity test are given in Hedges and Olkin (1985) and Rosenthal and Rubin (1982). This test statistic has an asymptotic chi-square distribution with $k - 1$ degrees of freedom.

If the results from the k studies are consistent with the model of a single common effect size, one can test whether the value of that single effect (δ) differs from zero. (If results are not consistent, one tests the value of the average effect). Thus the second hypothesis tested is the same hypothesis shown in Model 1. Hedges and Olkin (1985) noted that either z tests or confidence intervals can be used to test hypotheses about values of δ.

If the studies do not share a common population effect size, further analyses may be applied. One approach for further study of effect-size variability is to use a categorical model, analogous to the fixed-effects analysis of variance, to examine differences in results between *a priori* classes or categories of studies. Alternately, fixed-effects regression models allow the examination of fixed continous or categorical predictor variables, and random-effects or mixed models for variation in effect sizes (e.g., Hedges, 1983; Raudenbush & Bryk, 1985) can also be applied.

In summary, both effect-size analyses and modified tests based on vote-counts can test two kinds of meta-analysis hypotheses. They ask whether the effect sizes are consistent and they address the question of whether the common or average effect size is equal to zero. The tests of combined significance, however, can only test the composite null hypothesis that all the effect sizes are equal *and* equal to zero.

Statistical Properties

Bias

The maximum likelihood estimator for modified vote-counts proposed by Hedges (1986) has an asymptotic normal distribution with a mean of δ as k tends to infinity. Therefore, $\hat{\delta}$ is asymptotically unbiased.

The usual sample estimate of δ_i proposed by Glass (1976) is

$$g_i = (\bar{Y}_{1i} - \bar{Y}_{2i})/S_i,$$

where $\bar{Y}_{1i}$ and $\bar{Y}_{2i}$ are the two sample means in the ith study and S_i is the pooled within-groups sample standard deviation in the ith study. Hedges (1981) noted that g_i is a biased estimate of the population effect size δ_i, and provided an unbiased estimator $d_i = c(m_i)g_i$, where $c(m) = 1-3/(4m-1)$ and m is the degrees of freedom for the two-sample t test, the total sample size minus two. We will apply d_i in our effect-size analysis below.

Hedges and Olkin (1985) note further that weighted average estimates of δ (computed across studies) are only slightly biased due to the stochastic component in the weights (the inverse variances). This bias is minimal when the δ_i values are small and when samples are large.

Efficiency

Both maximum likelihood analyses based on vote-counts and effect-size analyses provide estimates for population effect sizes. Are both estimates equally good at estimating the population effect sizes? One aspect of this question is, are they equally efficient? Hedges (1986) indicated that the relative efficiency of the estimate of effect size based on vote counts compared with that of the parametric estimate based on sample effect sizes has a maximum at $\delta = 0$, independent of the sample size. This maximum efficiency is $2/\pi$ or 63.7%.

Thus, the estimate of population effect size based on vote-counts is less efficient than the parametric estimate. Its relative efficiency

decreases as δ moves away from zero and decreases more rapidly when n, an assumed uniform sample size in each study, is larger. Hedges and Olkin also pointed out that when δ and n increase, the parametric estimators are more likely to have the same sign and therefore relatively less of the information about effect size is contained in the signs of the estimate.

Statistical Power

Hedges and Olkin (1980) noted that if the average power of the studies is smaller than the cutoff criterion for the traditional vote-counting method, the statistical power of detecting a nonzero effect tends to zero as the number of studies increases. Thus even if all additional studies represent a nonzero effects, the likelihood of detecting that is decreased. Hedges and Olkin therefore suggested that the traditional vote-counting method may often be misleading.

Research has been done on the power of tests of combined significance (e.g., Koziol & Perlman, 1978). Becker (1985) indicated that only a few generalizations are possible. For example, the Tippett and Fisher tests are most powerful at detecting single large effects. Becker also noted that Stouffer's test has high power to detect small deviations from zero that are consistent across populations.

Little research has been done on the power of tests of homogeneity from effect-size analyses. Thus, the comparison of the statistical power among the three review methods is not available.

However, some comments may be made. There are two types of hypothesis testing involved with the three review methods: (1) tests of one composite hypothesis, and (2) tests done by considering the composite hypothesis through a two-step hypothesis test. The statistical power of the two types of hypothesis test will differ. The issue of Type I error also will complicate comparisons of the two types of hypothesis tests. The hypothesis-wise Type I error rate for a pair of tests will be twice that of

the single test unless it is reduced a priori, which in turn will affect the estimated power for the pair of tests. Such considerations would need to be made if power for effect-size analyses and combined significance tests were to be compared.

Examples

Data Set

Data for the example were drawn from 30 reports on gender differences in science achievement (Becker & Chang, 1986). These reports represent the published articles and ERIC documents from previous reviews by Steinkamp and Maehr (1983, 1984). Results from dissertations and test manuals were not included in our analyses.

Twenty-nine independent effect sizes were extracted from 23 of the 30 reports. Seven studies reported only the sign (direction) of the effect size, and twelve other sources cited by Steinkamp and Maehr had insufficient data for determining either the effect size or the direction of the effect size, or had reported on subjects who were also studied in another report.

Example 1: Overall Analysis

Our first example is based on an overall analysis of the study outcomes. No attempt is made to examine variation in study outcomes.

Traditional vote count. For the 29 samples for which the significance of the outcome could be determined, no sample produced significantly higher female achievement, nine samples (or roughly one third) showed significantly higher science achievement for males, and 20 samples showed no differences in science achievement between the sexes. (An overall α level of 0.05 was obtained by requiring results to be significant at the 0.025 level in both tails.) The winning category was the "no gender difference" outcome, since the bulk of the samples showed this result. Thus the conclusion from the

traditional vote count was that males and females do not differ on average science achievement.

Modified vote-counting methods. The vote-counting estimate of δ was obtained from all studies: those with effect sizes as well as those with only the directional results. Hedges' (1986) maximum likelihood vote-counting estimate of δ based on all 36 samples was 0.055 with a standard error of 0.025. A z test indicates that δ differs from zero ($z = 2.2$, $p < 0.02$) at the 0.05 level.

Hedges' likelihood ratio statistic provides a test of the overall homogeneity of results on the basis of the count data. The maximum value of the likelihood was -22.32, which gives a homogeneity chi-square value of $-2(-22.32) = 44.64$. The probability of obtaining a value of 44.64 or more is greater than .25 (as compared to percentage points of the chi-square distribution with 35 degrees of freedom). Thus on the basis of the vote-count data the results appeared homogeneous.

Combined significance tests. The directional probabilities summarized were for tests of the null hypothesis of no gender differences versus the alternative that males are superior to females on science achievement. That is, smaller sample probabilities were associated with samples in which males outscored females.

The Stouffer test value of 6.25 was highly significant when compared to a table of standard normal critical values. The Fisher test with a statistic of 192.05 was also significant when compared to critical points of the chi-square distribution with 58 (i.e., $2k$) degrees of freedom. Thus both combined significance tests rejected the null Model 3, and we conclude that at least one of the populations studied shows a significant male advantage on science achievement.

Effect-Size Analysis. Hedges' (1982) test of homogeneity of effect sizes was computed first to examine whether the results were reasonably

similar for the 29 samples providing effect sizes. The value of the overall homogeneity test statistic was 101.13, which was significant when compared to a chi-square with 28 degrees of freedom ($p < .001$). According to the effect-size analysis these studies do not appear to share a single common population value for the size of the gender difference.

The weighted average effect size for the 29 samples was 0.16 (with a standard error of 0.02), indicating slightly less than one sixth of a standard deviation advantage on science achievement measures for male students. This average effect differed significantly from zero at the .05 level (as indicated by a z test or confidence interval). However, because the homogeneity test indicated that the samples did not share a common effect size, the interpretation of the average effect size across the samples is not straightforward. Some samples can be expected to have a true gender difference larger than 0.16 standard deviations, while others may be from populations with smaller gender differences.

Summary. In the overall analysis we have applied four synthesis methods and addressed two hypotheses. Table 1 summarizes the results.

Table 1

Overall Hypotheses Examined by Four Research Synthesis Methods

	Average Effect of Zero	Homogeneity of Effect Size
Method	Hypothesis Supported?	
Traditional Vote-count	yes	not applicable
Modified Vote-count	no	yes
Tests of Combined Significance	cannot tell	cannot tell
Effect-Size Analysis	no	no

The traditional vote count indicated no gender difference, while providing no information regarding the consistency of gender differences.

The modified vote-count analyses suggested that the results of the studies of science achievement agree well, and furthermore that males and females do not differ significantly on mean science achievement. Either one or both hypotheses tested by combined significance are supported but without further information it is impossible to know which.

Finally, the effect-size analyses suggested that the gender differences were not consistent and that, on average, males outscored females on science achievement measures. Thus this result does not contradict the combined significance findings though it disagrees with the traditional vote-count homogeneity test.

Example 2: Further Data Analysis

Our second example is based on the hypothesis that gender differences may relate to the science content being examined.

Categorical analysis of effect sizes. Hedges' (1982) analogue to analysis of variance was used to further examine the variation in gender differences according to an a priori grouping of studies by the subject-matter content of the achievement measure. Table 2 shows the results of the categorical analysis according to subject matter.

Table 2

Subject-Matter Differences Among Effect Sizes

Source	df	Test of Homogeneity	p value	Mean effect-size estimate (SE)
Total	28	101.13	<.001	0.16 (0.02)*
Between subject-matter groups	4	51.14	<.001	
Within groups	24	49.99	<.001	
General science	10	30.68	<.005	0.07 (0.05)
Biology	5	4.49	ns	0.14 (0.04)*
Chemistry	0	0.00	-	-0.12 (0.06)
Geology	4	6.49	ns	0.10 (0.06)
Physics	5	8.33	ns	0.35 (0.04)*

The effect sizes within each subject-matter group were consistent or homogeneous with the exception of those from studies of general science. Mean effect sizes can be interpreted as common population effects for the five consistent subject-matter subgroups. Mean gender differences for studies of biology and physics both showed significant advantages for males, of 0.14 standard deviations for biology (with a standard error of 0.04), and of 0.35 standard deviation units for physics (SE = 0.03). The average gender effect for biology achievement also differed significantly from that for physics ($\underline{z} = 4.20$, $\underline{p} < .001$). Males and females performed equally on tests of both geology/earth sciences and chemistry (though the single study of chemistry provides little information for generalizations).

On average the gender difference for studies classed as general science did not differ from zero. However, because of the heterogeneity within the general science studies it is likely that some studies are of populations with effects that are nonzero.

Combined significance tests for subject-matter subgroups. To further illustrate the differences between the meta-analysis approaches the combined significance tests were applied to results within the subgroups of studies (except for the single study of chemistry). The tests were expected to indicate rejection of the composite hypothesis that all δ_i values equal zero for studies of biology and physics, because within these consistent groups the gender differences deviated significantly from zero. In fact both combined significance tests rejected the composite null hypothesis Model 3 for studies of biology ($\underline{z} = 3.20$; $\chi^2 = 30.91$, $\underline{df} = 12$; $\underline{p} < 0.002$) and physics ($\underline{z} = 7.20$; $\chi^2 = 97.56$, $\underline{df} = 12$; $\underline{p} < 0.00007$), and therefore agreed with the results from the effect-size analysis.

Effect-size analyses for studies of geology or earth sciences did not reject the hypothesis of homogeneity and also retained the hypothesis that δ

= 0 within the group. Thus the composite hypothesis that all effects equal zero for studies of geology is intuitively expected also to be retained on the basis of the combined significance tests. However, both combined significance tests were significant for this subgroup of studies. (The summaries were significant at the 0.02 level, $\underline{z} = 2.52$; $\chi^2 = 20.95$ $\underline{df} = 10$). Thus, the composite null hypothesis was rejected by the combined significance tests, and the conclusion was that at least one population studied showed a male advantage on geology or earth science achievement. This contradicted the effect-size analysis conclusion.

Further, the effect-size analyses indicated that the population effects for studies of general science were heterogeneous, though on average approximately zero. In order to agree with the effect-size analyses, the combined probability analyses would need to reject the null hypothesis of all effects equaling zero for the general-science studies. In fact, while the Fisher statistic was significant for the eleven studies of general science ($\chi^2 = 42.57$, $\underline{df} = 22$, $\underline{p} < 0.02$) the Stouffer test was not ($\underline{z} = 1.53$, $\underline{p} = 0.06$). Here there is even disagreement among the combined probability analyses concerning gender differences in the general-science studies. Analyses support both the presence and absence of gender differences on general science.

Discussion

The comparisons made in the previous section do not enable us to determine which (if any) of the competing analyses has given the correct answer because the comparisons are based on sample data. However, insights can be gained by comparing the outcomes and conclusions for the analyses in light of the strengths and weaknesses of each.

Example 1

The first comparison which can be made is between the overall analyses

(tests of homogeneity and estimates of effect size) based on vote counting and on effect-size analysis. It is not surprising that the traditional vote count, with its low power did not support the finding of gender differences in science achievement. Becker (1986) noted a similar pattern of conclusions for studies of gender differences in social influence.

The vote-counting estimate of effect size from the analysis of all 36 study results was 0.055, or about one third the value of the parametric estimate. However, the value differed significantly from zero (at the .05 level), thus supporting the same general inference as the parametric test of $\delta = 0$.

To address the question of whether the reduced value of the vote-count estimate resulted from the addition of studies without effect sizes, we computed vote-count estimates of δ for the studies with and without effect sizes. The values for the two sets of studies are quite similar (0.06 for the 29 studies with effect sizes and 0.04 for the seven studies reporting only directions of results) and both are smaller than the parametric estimate. This suggests that other factors have caused the vote-count estimate to be smaller than the parametric estimate.

Unlike average effect sizes which are based on the sample mean differences from each study, the estimate of effect size from vote-counting ($\hat{\delta}$) is based on the directions of the differences between two groups across the whole set of studies. Thus information about the sizes of the mean differences in the samples under review is unavailable to the vote-count estimate which is especially crucial when δ_i's are large.

Hedges and Olkin (1985) noted that when all study results have the same direction, there is no unique estimate of effect size and when <u>most</u> of the study effect sizes have the same direction, the estimate of δ is less accurate than when there is greater balance in the signs of the results. In

the example data large proportions of the results were positive: Twenty-three of the 29 results (79%) of studies with effect sizes were positive (indicating higher male achievement scores) and six of the seven studies reporting only directions of results noted that males had outscored females.

Furthermore, over half of the calculable effect sizes in this example were larger than 0.15 standard deviation units. If the calculated effect sizes represent reasonably well the true size of the gender difference in science achievement, the vote-counting estimate is considerably less efficient than Glass's effect size estimate for this data.

The overall test-of-homogeneity results from the vote-count analysis indicated that the study results were consistent, while those of the effect-size analysis indicated considerable lack of agreement among the results. This resulted in large part again from the reliance of the vote-counting tests on only the sign (as was true in this data), the magnitudes of population effect sizes may still vary, and the differences can not be detected by the likelihood ratio test.

The results of the tests of combined significance agreed with the overall effect-size tests of homogeneity and of average effect magnitude, though they did not provide the same information. Effect-size homogeneity tests indicated that all studies did not share a common population outcome. One could not distinguish whether rejection of the null Model 3 on the basis of combined significance tests resulted from heterogeneity of the study results or from deviation of any of the effect sizes from zero. The differences in the conclusions for these two kinds of test become more obvious when comparing the significance tests with further effect-size analyses.

Example 2

Further comparisons between the methodologies can be made by considering the results of the categorical analyses of studies according to

subject matter. First we note that vote-count estimates and tests could not be computed for the subject-matter subgroups of studies because of the similarity of outcomes within the groups. If the population values of parameters within subgroups of studies are similar and different from zero this problem will be most likely to occur. Also, the problem is more likely to occur when subgroups of studies are small (even if the population parameters for the studies are near zero).

When further analyses of effect size were applied, three different patterns of outcomes were identified. In studies of biology and physics the gender differences were relatively homogeneous and differed significantly from zero. Gender differences for studies of geology/earth sciences were also homogeneous, but the common gender difference was essentially zero. Finally the results of studies of general science were not consistent with each other and their average value was near zero.

The three patterns of effect sizes were associated with three very different substantive conclusions about the nature of gender and science achievement. However, the oucomes of tests of combined significance for these subsets of studies were almost identical. It was expected that the combined significance tests would be significant for the studies with nonzero common effects (i.e., those of biology and physics) or heterogeneous effects (i.e., the general-science studies). However, the tests also rejected the null hypothesis of equal effects for the studies of geology/earth science. More so, one of the combined significance tests, namely Stouffer's test, disagreed with Fisher's test of combined siginificance, and failed to reject the null hypothesis of equal effects for the general-science group.

Examination of power values for the combined significance tests indicates that Fisher's combined significance tests (used in our analysis)

is very likely to reject the null hypothesis for even small single deviations. Also Stouffer's test has high power to detect small deviations from zero that are consistent across populations (Becker, 1985). These power differences may explain the "conflicting" results from these two tests of combined significance. For our data set, the Fisher test may have been more powerful than Stouffer's test. The difference between the results of the combined significance and the effect-size analyses may have resulted from the lower power of the effect-size analysis. However, more study on power of homogeneity tests needs to be done to make stronger conclusions on this matter.

Summary and Recommendations

Effect-size analyses appear more informative for the study of single outcome constructs than the other methodologies examined because they enable reviewers to test hypotheses about specific values for the population effect magnitudes (even on the basis of single sample effects), and to formulate models for predicting effect sizes from study features. Tests of combined significance are omnibus tests and therefore by definition less informative. Applying combined significance tests with effect-size analyses also does not provide more information, while additionally increasing the chance that the reviewer will make a Type I error because the analyses are not independent.

Thus tests of combined significance are not preferable to effect-size analyses. They might be useful when effect sizes can not be analyzed, for example, when effect sizes can not be obtained but probability values are available (which is unusual), or when studies present tests of different parameters or constructs which can not be transformed onto a common scale. Tests of combined significance can appropriately be applied in these situations but valid interpretations of their results will still be very broad.

If only the signs of the outcomes are presented, vote-counting methods are the only applicable method, but reviewers should realize the limitations of the procedures. Information about magnitude of effect simply is not available to the vote-count procedures. Thus vote-count effect-size estimates will have relatively larger standard errors than would be obtained for parametric estimators, especially when the true value is far from zero.

A few conclusions may be drawn also about gender differences in science achievement. Gender differences for all subject-matter groups, except for studies of general science, were consistent, and in many areas no significant differences were found. Even the significant differences were all less than one half of a standard deviation: in physics males outscored females on the average by about one-third of a standard deviation while in biology they outperformed females by about one sixth of a standard deviation. These results (though tentative) suggest that care should be taken to carefully distinguish between subject-matter areas when discussing or researching science achievement and gender.

References

Becker, B. J. (1985). Applying tests of combined significance in meta-analysis. Paper under review for Psychological Bulletin.

Becker, B. J. (1986). Influence again: An examination of reviews and studies of gender differences in social influence. In J.S.Hyde & M.C. Linn, (Eds.). The psychology of gender: Advances through meta-analysis. Baltimore: Johns Hopkins.

Becker, B. J. (in press). Applying tests of combined significance in meta-analysis. Psychological Bulletin.

Becker, B. J. & Chang, L. (1986). The measurement of science achievement and its role in gender differences. Paper presented at the annual meeting of the American Educational Research Association.

Fisher, R. A. (1932). Statistical methods for research workers. (4th ed.). London: Oliver & Boyd.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.

Hedges, L. V. (1982). Fitting categorical model to effect size data. Journal of Educational Statistics, 7, 245-270.

Hedges, L. V. (1986). Estimating Effect Size from Vote Counts or Box-score Data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hedges, L. V. & Olkin, I. (1980). Vote-counting methods in research synthesis. Psychological Bulletin, 88, 359-369.

Hedges, L. V. & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, Florida: Academic Press.

Koziol, J. A. & Perlman, M. D. (1978). Combining independent chi-square tests. Journal of the American Statistical Association, 73, 753-763.

Kraemer, H. C. (1983). Theory of Estimation and Testing of Effect Sizes: Use in Meta-Analysis. Journal of Educational Statistics, 2, 93-102.

Raudenbush, S. W., Bryk, A. S. (1985). Empirical Bayes meta-analysis. Journal of Educational Statistics, 10, 75-98.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Rosenthal, R. & Rubin, D. B. (1982). Comparing effect sizes of independent studies. Psychological Bulletin, 92, 500-504.

Steinkamp, M. W. & Maehr, M. L. (1983). Affect, ability, and science achievemeny: A quantitative synthesis of correlational research. Review of Educational Research, 53, 369-396.

Steinkamp, M. W. & Maehr, M. L. (1984). Gender differences in motivational orientations toward achievement in school science: A quantitative synthesis. Journal of the American Educational Research Association, 21, 39-59.

Tippet, L. H. C. (1931). The methods of statistics. (1st ed.). London: Williams & Norgate.